Developing a Simple Question-Answer System

Northwestern University IEMS 308
Jake Atlas

Executive Summary

This project serves to create a simple question and answer (QA) system that is capable of answering three question types. By analyzing a corpus of articles scraped from the web, the developed QA system can conduct searches to identify segments of text that likely contain answers to the question that has been posed. Such a system is incredibly useful for determining answers to simple questions as well as for understanding what information is actually contained in the corpus of documents.

Natural language processing and text analytics are analytical techniques that serve to facilitate and perform a variety of searches within textual data. Also generally referred to as text mining, text analytics is an integral part in developing a system capable of parsing questions, searching for answers, and outputting the best answer or set of answers. Using text mining techniques to develop a QA system has practical value in that it enables factoids to be extracted from large datasets quite quickly.

After organizing the text data and indexing it using Elasticsearch, a distributed search and analytics engine, a system was developed that parses a question and subsequently repeatedly generates subsets of the corpus, homing in on candidate answers. Based on patterns in the data and the type of question, an answer is then selected from the candidates and outputted to the user.

The results proved to be favorable, in that the answers provided for the test questions are reasonable and take no more than thirty seconds processing time. There is, however, room for improvement, as there is a limit to the questions that can be answered with this particular corpus. This QA system also focuses on three particular types of questions, and generalizing it such that further questions can be answered would be a significant improvement.

Problem Statement

The supplied corpus of articles provides a wealth of information about contemporary happenings in the business sphere. My job was to apply text mining procedures to this data to create a question and answer system that is capable of answering the following questions:

1. Which companies went bankrupt in monthX of yearY?
2. What affects GDP? What percentage drop or increase is associated with this property?
3. Who is the CEO of companyX?

Methodology

The first step in creating a question and answer system that leverages the information in the corpus is to install one of the professional packages for indexing the data. For the purpose of this project, Elasticsearch was chosen, and the Python API was utilized. Therefore, the first step is to ensure that Elasticsearch is running properly.

Once it has been verified that Elasticsearch is functioning latently, the corpus, which is currently stored as 730 separate text files in a locally stored folder, can be indexed using Elasticsearch. In this step, each document is indexed separately, such that searches retrieve full documents. Further steps will re-index relevant subsets of these documents in order to narrow the search for answers. Figure 1 below depicts the indexing of the 730 articles in the corpus and displays the relevant information that Elasticsearch uses to identify and distinguish the documents. Accessing the articles is shown over a blue background, and indexing them is shown over a green background.

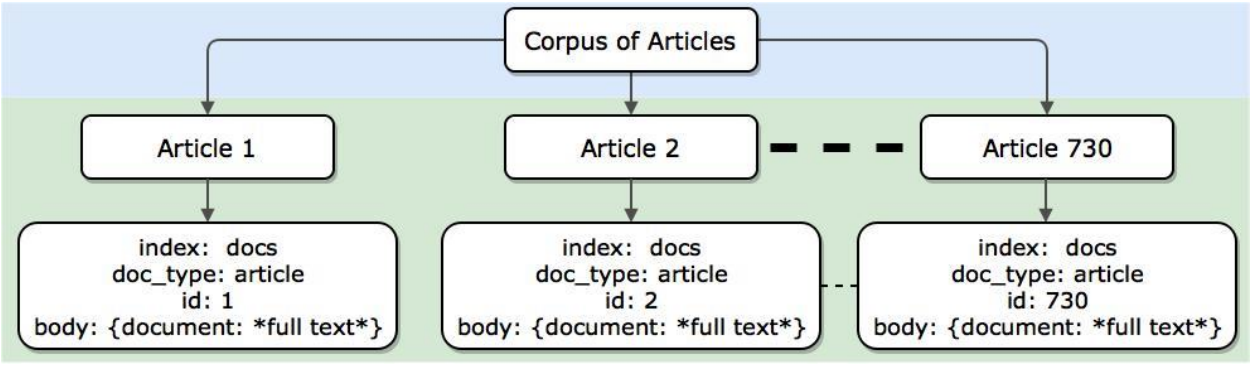

Figure 1

Once the articles have been indexed, sentence segmentation is performed on the original corpus for future use. By separating each article into its constituent sentences, it becomes possible to narrow the search for an answer to the sentence level once candidate documents have been identified.

The system now prompts the user to ask a question. The question is word-tokenized, stop words are removed, and the question is classified as one of the three types listed above in the problem statement. The remaining words are the question keywords, which are vital in searching for relevant information in the text. Based on the question classification and remaining keywords, a search is conducted to identify documents that may contain an answer to the question that the user posed. The documents are selected based on term frequency – inverse document frequency (tf-idf) scores.

From this set of documents, the individual sentences are then indexed. Figure 2, which appears at the top of the next page, provides a visualization of this process, highlighting the indexing of the sentences in green. Steps that have already been completed are highlighted in blue. Note that while this diagram shows the indexing of the sentences from only one article with ID n, each article containing a potential answer is indexed in this same fashion.

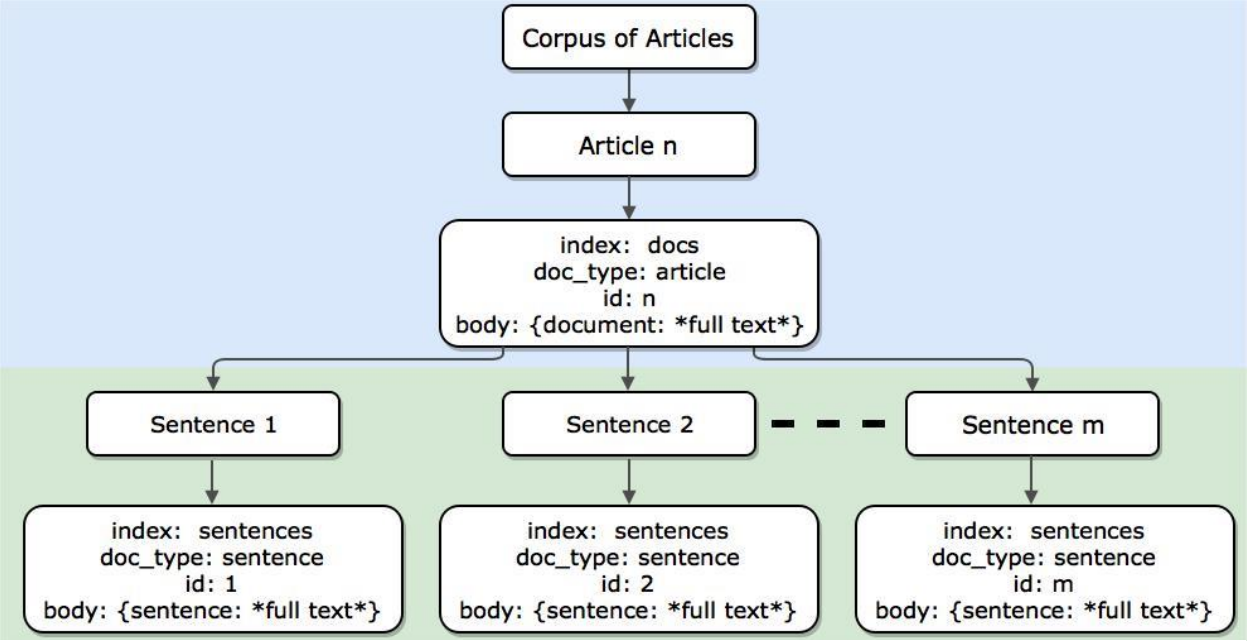

Figure 2

With the sentences indexed, a search can now be performed on the sentences, utilizing the same keywords as the search for full articles. As with the search for relevant articles, the sentences are scored using tf-idf and ranked accordingly. The highest-ranking sentences are then searched for answers by leveraging the classification of the question. Depending on which question was asked, different regular expressions are used to search for answers within sentences.

In some cases, before regular expressions are used to search for the answer, further subsets of the sentences are created based on the location of keywords. This effectively applies a heavy weighting to those particular keywords. For example, in asking about bankruptcy, segments of text around the word “bankrupt” are used as subsets. This ensures that answers are relevant to the question asked. Finally, the selected answer or compiled answers are outputted to the user.

Analysis

The entire analysis is focused on the quality of the answers provided. The analysis will be broken into three subsections, one for each of the three types of questions that this QA system is designed to answer. Throughout the analysis, there are examples from actual questions posed to the QA system. For a complete list of examples, see Appendix I on page 7.

Question 1: Which companies went bankrupt in monthX of yearY?

The answers to this question are good in that they are, in fact, companies that have gone bankrupt. The answers are not perfect though, as they often seem to not correspond to the month and year specified in the question statement. This is likely due to the infrequency with which the articles reference bankruptcies using month and year. Further examination

shows that, for almost all combinations of month and year, there are no articles that contain a reference to bankruptcy with both month and year. As a consequence of this, the answers displayed are often very similar, regardless of the month and year that have been specified. This is not a shortcoming of the QA system structure, and is instead demonstrates the important point that a QA system is only as smart as its corpus. The system is unable to answer questions unless the information is contained in the corpus.

Question 2: What affects GDP? What percentage drop or increase is associated with this property?

The answers to this question are quite good in that all suggested factors contributing to GDP do have a connection to GDP. Where the answers to this question falter is in reporting the associated percentage drop or increase in GDP. While all follow-up questions regarding the percentages are successfully answered, resulting in a percent value being reported back to the user, these percentages are not actually meaningful. For example, to say that unemployment causes a 4.2% change in GDP tells only a situational truth. For the particular increase or decrease in unemployment discussed in the corpus, there was a 4.2% change in GDP; however, not every change in unemployment will result in the same percentage response in GDP. Generally speaking, though, the constructed QA system does a very good job answering this question, especially considering that it is not quite a factoid question. Posing this question to WolframAlpha, a highly regarded QA system, absolutely stymies it; instead of providing an answer, the system simply defines the word “affect.” Given the relative performance of the QA system I have developed over WolframAlpha, the QA system developed for this project appears quite capable.

Question 3: Who is the CEO of CompanyX?

The answers to this question are correct, which is unsurprising since it is a simpler, factoid question. While the ability of the developed QA system is reliant on the corpus for answers, the articles contained in the corpus do make references to the CEOs of America’s larger and more commonly spoken about companies. As a result, the QA system is capable of quickly determining the CEO of companies such as Amazon, Apple, and Facebook, among others. Of course, since the corpus contains articles written a few years ago, there are times where the reported CEO of a company is no longer the CEO, as in the case of Larry Page and Google. This again highlights the reliance of a QA system on its corpus of documents. However, broadly speaking, the developed system is successful at determining company CEOs.

Conclusions

The key insights from the QA system developed to extract information from the corpus of scraped business-related articles are:

A question-answer system is only as smart as its corpus. There is simply no feasible way for the QA system to provide information that is not contained in the corpus. Consequently, it is

essential that the architect of a QA system understands the type of information contained in the corpus.

It is extremely challenging to extract answers to complex questions. Non-factoid style questions are much harder to answer due to the fact that answers may change or have multiple parts and also because the question often does not contain the same caliber keywords to use in searching as in factoid questions. For example, a question such as "What affects GDP?" is difficult to turn into keywords without utilizing a word bank such as WordNet. The word "affects" would need to be supplemented by words such as: change, increase, decrease, shrank, dropped, rose, fell, and others in order to fully capture the variety of terms that may be used interchangeably for discussing the effect of some property on GDP.

These key insights are validated both by the answers generated for the questions that were posed as well as by professional QA systems that are currently used. I am therefore confident that despite the shortcomings of the developed QA system, it is overall quite successful. I am also confident that adding more relevant information to the corpus would enable improvements to the model.

Next Steps

While the QA system successfully answers the three questions posed, it is somewhat limited to those three question types. In order to design a more robust model capable of answering a variety of questions, it makes sense to employ a connection to WordNet in order to generalize questions. This enables more effective searching; as demonstrated by this project, there are often many ways to phrase a question and answer, and a robust model should consider these potential variations.

It would also be interesting to observe the degree of improvement of the QA system should more documents be added to the corpus. In answering the question regarding bankruptcy in particular, the effects of missing information were highlighted. It would be a worthwhile experiment to see whether this question and similar questions can be answered more specifically with the addition of documents to the corpus. In this case, the documents would have to contain information regarding bankruptcy, as it has been made clear that questions cannot be answered by a corpus that does not contain the answer.

With these advancements made, it makes sense to compare the new, higher-powered QA system to other already well-established QA systems, like WolframAlpha, which also powers Siri. We have seen already that WolframAlpha fails to answer such questions as "What affects GDP?" so with further advancements made on this model, a comparison of strengths and weaknesses would be quite interesting.

Appendix 1 – Sample Questions and Answers

Question 1:

1. What companies went bankrupt in April of 2011?
    a. ['Allied Crude', 'Estar']
2. Which companies went bankrupt in September of 2005?
    a. ['Solyndra', 'Estar', 'Glaziev']
3. Which companies went bankrupt in June of 2010?
    a. ['Allied Crude', 'Estar']

Question 2:

What affects GDP?

a. ['PMI', 'unemployment', 'austerity']
    i. What percentage drop or increase is associated with PMI?
        a. 3.2 percent
    ii. What percentage drop or increase is associated with unemployment?
        a. 4.2 percent
    iii. What percentage drop or increase is associated with austerity?
        a. 18 %

Question 3:

1. Who is the CEO of Amazon?
    a. Jeff Bezos
2. Who is the CEO of IBM?
    a. Ginni Rometty
3. Who is the CEO of Facebook?
    a. Mark Zuckerberg